[RCG] Revista Catalana de Geografia
Revista digital de geografia, cartografia i ciències de la Terra

Autor/s: Guerrero Nieto, M. , García Rodríguez, M. J. , Urrutia Zambrana, A. , Bernabé Poveda, M. Á.
Títol: Propuesta para la integración de expresiones temporales procedentes de patrimonio documental en un SIG
Publicat a: Revista Catalana de Geografia IV època / volum XV / núm. 40 / juliol 2010
Font: -
URL: http://www.rcg.cat/articles.php?id=179

PROPUESTA PARA LA INTEGRACIÓN DE EXPRESIONES TEMPORALES PROCEDENTES DE PATRIMONIO DOCUMENTAL EN UN SIG

Marta Guerrero Nieto, María José García Rodríguez, Adolfo Urrutia Zambrana, Miguel Ángel Bernabé Poveda
Grupo Mercator - Universidad Politécnica de Madrid

Resumen

En este artículo se propone una metodología para la integración de un corpus histórico procedente de textos originales en un Sistema de Información Geográfica (SIG), donde destaca la incorporación de los aspectos temporales del corpus en la base de datos de un SIG. Para ello, se ha utilizado el lenguaje de marcado TimeML que facilita la comunicación entre los SIG y la información procedente de patrimonio cultural. Al amparo de este campo de estudio, el presente trabajo tiene dos objetivos fundamentales: por un lado, la identificación y normalización de expresiones temporales, y por otro lado, la incorporación de la variable temporal extraída de corpus históricos en los SIG.

Introducción

Esta investigación se enmarca dentro del contexto del proyecto DynCoopNet (Dynamic Complexity of Cooperation-Based Self-Organizing Commercial Networks in the First Global Age), que tiene entre sus propósitos indagar en la dinámica de las redes comerciales de cooperación que se establecieron durante la Primera Edad Global (1400-1800), mediante la promoción del uso y desarrollo de los Sistemas de Información Geográfica (SIG) en aplicaciones con componentes históricos en el campo de las Ciencias Sociales y Humanidades. Además persigue asímismo abordar estudios de confrontación, revisión y reconocimiento de patrones de fenómenos o actividades humanas y profundizando en la visualización de narrativas históricas [1].

El corpus utilizado procede de una selección de cartas del comerciante español Simón Ruiz fechadas en el s. XVI. Estos datos forman parte del patrimonio cultural desde dos aspectos: externo, ya que proceden del Archivo Histórico Provincial de Valladolid, que está formado por más de 56.000 cartas comerciales datadas entre los años 1553 y 1630; y desde el contenido interno, pues se relatan las redes comerciales que germinaron en Castilla, siendo el eje del comercio Medina del Campo, y se relacionaron con las principales ciudades europeas, extendiéndose hacia los territorios recién descubiertos de América.

Para llevar a cabo esta investigación se propone el uso del campo Procesamiento del Lenguaje Natural (PLN), donde se incluye la anotación semántica, en la aplicación al contenido patrimonial. Este campo de las Humanidades y Ciencias Sociales revela un auge en las investigaciones desarrolladas en los últimos años con respecto al uso del SIG [2][3][4]. Para poder integrarse en un SIG, el corpus elegido debe acomodarse a unas series de especificaciones, ya que posee una naturaleza singular, pues se trata de un compendio de textos escritos en lenguaje natural, y una estructura distinta a la que se requiere por parte de la herramienta geográfica. Teniendo esto en cuenta, se ha acudido al modelado de los lenguajes de marcado. Dado que los SIG son capaces de representar la espacialidad del patrimonio cultural, se pretende la extensión para la componente temporal. Para ello, se utilizará el lenguaje de marcado temporal TimeML [5].

Procesamiento de Lenguaje Natural (PLN) para la extracción de información temporal

El PLN es una subdisciplina de la Inteligencia Artificial que tiene como propósito el modelado y procesamiento computacional del lenguaje humano. En la temática que nos ocupa, el procesamiento de la información temporal ha despertado un gran interés en la comunidad científica, prueba de ello son los numerosos workshops celebrados en diferentes áreas como en la creación de herramientas de extracción y análisis temporal (TERCAS [6], TANGO [7], DAGSTUHL [8], MUC [9]); los lenguajes de anotación semántica temporal (TIDES [10], TimeML [11]); los sistemas de anotación (TERSEO [8]) y en diferentes talleres de evaluación (TERN [12], TemEval [13]).

Con la incorporación del campo del PLN para el análisis de la temporalidad, se pretende utilizar las expresiones temporales lingüísticas del documento, ya sean estas explícitas o implícitas, para así poder situar en el tiempo los acontecimientos descritos y cuantificarlos, que respondan por ejemplo a las siguientes preguntas: ¿cuándo ocurren los eventos?, ¿cuánto duran?, ¿en qué periodo se dan?, etc. Normalmente las bibliotecas digitales, repositorios de datos digitales o catálogos de datos digitales contienen una información temporal extraída de los metadatos del documento. Estos datos resultan claves para las consultas temporales, pero son insuficientes si se quiere consultar

acerca de la duración de acontecimientos u obtener otras fechas distintas a la fecha de publicación de dicho documento [14]. Esta carencia conduce a que el objetivo no sea únicamente el uso de los metadatos del documento sino la extracción de la información relevante contenida en el texto.

A continuación se presenta una breve descripción de la anotación lingüística, aplicación para el tratamiento del lenguaje natural, que incluye la anotación semántica temporal (TimeML).

Anotación lingüística

Como se ha comentado anteriormente, en este estudio se emplearán los lenguajes de marcado para la extracción de la información temporal de los acontecimientos históricos. Un documento en lenguaje natural puede ser marcado con el fin de identificar la información que se necesite; para ello se enriquece el texto con etiquetas. Con éstas, la estructura de un documento se hace explícita, otorgándole información adicional.

El metalenguaje utilizado habitualmente para marcado descriptivo o semántico es XML. Este metalenguaje es usado por la Lingüística Computacional para incrustar la etiqueta en el texto mediante la anotación lingüística que se le está añadiendo al documento. Las ventajas de la anotación lingüística son: en primer lugar, la posibilidad de utilizar con posterioridad técnicas de aprendizaje y creación de herramientas que lleven a cabo la etiquetación automática de corpus, por otro lado, la reutilización del corpus anotado se convierte en una fuente de investigación científica para la investigación y el desarrollo futuros.

Lenguaje de marcado temporal (TimeML)

Para la modelización, transporte y almacenamiento de la información temporal procedente de las cartas de Simón Ruiz se ha usado TimeML. Este lenguaje de marcado es una especificación lingüística para anotar eventos y expresiones de tiempo que ofrece una sistematización para la extracción y representación de información temporal, así como para el intercambio de información. Las propiedades más características de este lenguaje son: interpretación de las expresiones temporales, anotación del tiempo de los eventos y ordenación de los eventos con respecto a otros a través de un anclaje temporal. TimeML ha sido consolidado como estándar ISO (ISO WD 24617-1:2007) y es compatible con la forma ISO 8601 para el almacenamiento de las fechas.

Descripción y características del TimeML

Este lenguaje de marcado contiene tres etiquetas básicas: TIMEX3, EVENT, SIGNAL y tres subtipos de link: TLINK, ALINK y SLINK. Se procede a explicar brevemente cada una de las etiquetas:

- TIMEX3 se usa para marcar expresiones temporales: *15 de diciembre de 2009, ayer, a las 6 de la tarde, este sábado, el próximo año* .
- EVENT se usa es usada para marcar eventos mencionados de un texto: *ocurrir, creer, estudiar, empezar* .
- SIGNAL es usada para anotar señales temporales: *antes, después, durante* .
- TLINK se usa es usada para marcar las relaciones temporales: *Rosa trabaja desde las 5 de la tarde hasta las 12 de la noche*.
- ALINK se usa es usada para anotar las relaciones aspectuales: *Rosa terminará de trabajar hoy a las 11 de la noche* (el verbo *terminar* está mostrando una fase del evento).
- SLINK se usa es usada para anotar relaciones de modalidad o evidencialidad: *Rosa cree que no irá a trabajar mañana* (se marca la conjetura ante la realización del evento)

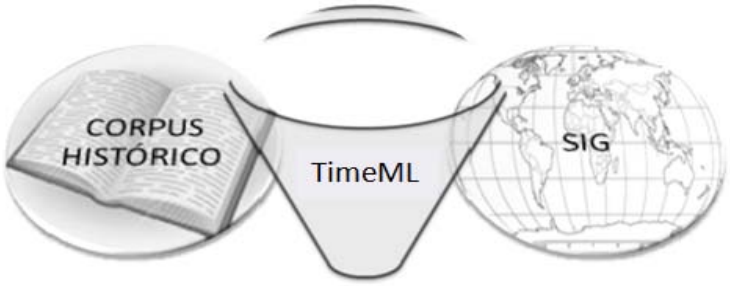
TimeML ofrece la posibilidad de expresar distintas granularidades. Posee cuatro tipos para la expresión de tiempo (TIMEX3):

- DATES se utiliza para expresiones que se refieren a un calendario: *a 14 de diciembre de 2009, el domingo pasado, ayer por la mañana* , etc.
- DAY TIMES es utilizada para una expresión temporal que es menor a un día: *esta noche, a las tres menos veinte* . La distinción entre estos dos tipos de tiempos tiene especial atención por la diferencia entre la granularidad de las expresiones.
- DURATION se usa para describir una duración en el tiempo: *durante cuatro años, hace dos días* .
- SET se usa para expresiones que se repiten en el tiempo: *tres veces a la semana, cada ocho días* .

Propuesta para la incorporación de la temporalidad en un SIG

El objetivo de esta propuesta es presentar una metodología para la incorporación de la componente temporal procedente de un corpus en un SIG. El procedimiento seguido se divide en dos fases:

- Identificación y normalización de las expresiones temporales del corpus.
- Incorporación de TimeML en una Geodatabase.



Identificación y normalización de las expresiones temporales

Para el reconocimiento y normalización de las expresiones temporales en lengua española no existe ninguna herramienta automática asociada a TimeML. Por este motivo, la prueba sobre el corpus histórico escrito anteriormente se ha realizado de forma manual. Al tratarse de un corpus histórico del español, ha sido necesaria la adaptación de la guía, ya descrita para la lengua inglesa, de acuerdo a la variedad lingüística de nuestro corpus. Es pertinente señalar que por el momento no ha habido trabajos enfocados en la especialización de correspondencia en castellano ni en XML ni en otro formato que incluya información temporal.

Tras la identificación de las expresiones temporales en castellano renacentista, el siguiente paso ha sido la normalización de estas expresiones siguiendo la definición de TimeML. A continuación se muestra un ejemplo del corpus de la normalización de estas expresiones temporales en lenguaje TimeML, donde aparecen algunos de los valores de la guía: TIMEX3, EVENT y TLINK.

"La de v.m. de 15 deste he recebido en este dia"

15
deste
vform="NONE" class="OCURRENCE" tense="PRESENT"
stem="RECIBIR"> **recebido**
anchorTimeID="tid11">**este dia**
28 de mayo de 1567
timeID="tid13" relatedToEventInstance="eid28"/>
lid="lid31"timeID="tid12"relatedToEventInstance="eid28"/>

Como puede observarse en el ejemplo, las expresiones lingüísticas utilizadas en las cartas pueden ser deícticas, esto es, necesitan del conocimiento del momento narrativo en el que se enmarcan para poder precisar el intervalo de tiempo comprendido por la expresión. Este lenguaje utilizado permite definir eventos y expresiones temporales con el fin de poder determinar el momento de ocurrencia y así poderlos situar en una línea de tiempo. Esto se consigue con el atributo *AnchorTime*, que se puede ver en el ejemplo, el cual permite establecer un eje temporal.

La etiqueta que marca las relaciones temporales es el TLINK, basándose éstas en las trece relaciones binarias del álgebra temporal de Allen [15]. Los TLINK representan las relaciones temporales existentes entre dos eventos, dos tiempos o entre un evento y un tiempo. En el ejemplo, el evento sería "he recebido" que va acompañado de dos expresiones temporales "de 15 desde" y "en este día". Las relaciones temporales entre estos tres elementos se marcan con la etiqueta TLINK de la manera que se puede ver en el ejemplo.

Con objeto de garantizar la consistencia de la estructura de los datos en todos los documentos se ha utilizado una DTD (*Document Type Definition*), una descripción del formato de los datos y la estructura del documento, de sus elementos y del anidamiento de las etiquetas.

Incorporación de TimeML como parte de una geodatabase

Una vez anotado el corpus en TimeML se continúa con la integración del texto en el Sistema de Información Geográfica. La DTD de TimeML proporciona una estructura estable y predecible, por lo que se podría diseñar una base de datos relacional para almacenar la información contenida en los atributos de cada uno de los elementos del archivo. Esta gramática (DTD) contiene todos los elementos, valores y atributos de los que está compuesto el TimeML.

Para poder automatizar el traspaso de información se requiere de la creación de un algoritmo de mapeo entre ambas estructuras (BD y Corpus), con el fin de poder guardar y extraer la información libremente. Tal herramienta podría ser implementada como un módulo interno del gestor de la base de datos o como un componente de software independiente [16]. Debido a lo anterior, las dos entidades (geodatabase y DTD) serían prácticamente idénticas, facilitando la introducción de la información. Así, el XML y la geodatabase se convierten en las dos caras del almacenamiento de las expresiones temporales.

Finalmente, con la información dentro de la geodatabase, la representación del corpus anotado dependerá de sus características y contenido. Encaminándose hacia el estudio y visualización de narrativas históricas.

Conclusión y trabajos futuros

Se ha representado el tratamiento de textos con información histórica y perteneciente al patrimonio cultural para su incorporación en un SIG a la vez que se ha establecido una metodología para el reconocimiento y la normalización de expresiones temporales procedentes de un corpus histórico, siguiendo las especificaciones del TimeML. Así mismo se han mencionado las ventajas de la utilización de este lenguaje de marcado, ya que posee un carácter

estándar, es compatible con la estructura de una base de datos, puede aplicarse a cualquier lengua y permite definir un tipo de expresiones temporales no especificado en otros lenguajes.

De igual modo se han descrito algunas de las limitaciones que se han observado para llevar a cabo la propuesta: (a) para lograr la representación de la información temporal etiquetada en el corpus, el SIG deberá tener una base de datos espacio-temporal que le permita almacenar y consultar la información proveniente del corpus; (b) existe una escasez de corpus etiquetados en TimeML para lenguas diferentes al inglés, lo que impide la utilización de técnicas de aprendizaje automático, motivando el uso de la anotación semiautomática y manual; (c) se necesita la adaptación del TimeML al castellano antiguo para facilitar la identificación de expresiones temporales en este tipo de textos.

Entre los trabajos futuros, se pretende conseguir el reconocimiento y normalización de expresiones temporales en corpora históricos del español más amplios, así como la integración de anotación lingüística temporal y espacial.

Referencias

- 1. Owens, J.B.: Visualizing the Past: Tools and Techniques for Understanding Historical Processes, (Working Papers), University of Richmond, Virginia, USA (2009)
- 2. Gregory, Ian N; Paul S. Ell.: Historical GIS: Technologies, Methodologies and Scholarship. Cambridge: Cambridge University Press (2007)
- 3. Knowles A.K.: Introducing Historical GIS, in Knowles A.K.(ed.) Past Time, Past Place: GIS for History, ESRI Press: Redlands, CA (2002)
- 4. Knowles, A. K., Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship. ESRI Press (2008)
- 5. Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G., Radev, D.: TimeML: Robust specification of event and temporal expressions in text. In: AAAI Spring Symposium on New Directions in Question-Answering (Working Papers), Stanford, CA, pp. 28--34 (2003)
- 6. Pustejovsky, J.: TERQAS: Time and Event Recognition for Question Answering Systems. ARDA Workshop, MITRE, Boston (2002).
  Available at http://www.timeml.org/site/terqas/index.html
- 7. TANGO (TimeML Annotation Grafhical Organizer),
  http://www.timeml.org/site/tango/index.html
- 8. Dagstuhl Seminar Proceedings. Annotating, Extracting and Reasoning about Time and Events,
  http://drops.dagstuhl.de/opus/volltexte/2005/313/
- 9. Advanced Research Projects Agency. Proceedings of the Sixth Message Understanding Conference (MUC-6) (1995). Software and Intelligent Systems Technology Office.
- 10. Ferro, L., Gerber, L., Mani, I., Sundheim, B., & Wilson, G.: TIDES 2005 Standard for the Annotation of Temporal Expressions. The MITRE Corporation (2005)
- 11. Saquete, E., Martínez-Barco, P., Muñoz, R., Negri, M., Speranza, M., Sprugnoli, R.: Automatic resolution rule assignment to multilingual Temporal Expressions using annotated corpora. In: Proceedings of the Thirteenth International Symposium on Temporal Representations and Reasoning, pp. 218--224 (2006)
- 12. DARPA TIDES (Translingual Information Detection, Extraction and Summarization). The TERN evaluation plan: Time Expression Recognition and Normalization. Working papers, TERN Evaluation Workshop (2004).
  Available at http://timex2.mitre.org/tern.html
- 13. Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., Pustejovsky, J.: SemEval-2007. Task 15: TempEval Temporal Relation Identification. In: Proceedings of SemEval 2007, 4th International Workshop on Semantic Evaluation, ACL, Prague, pp.75--80 (2007)
  Available at http://nlp.cs.swarthmore.edu/semeval/tasks/index.php
- 14. Llidó Escrivá, D. M.: Extracción y recuperación de la información temporal, Thesis Universidad Jaume I, Castellón, Spain (2002)
- 15. Allen, J. F.: Maintaining knowledge about temporal interval, Comunications of ACM, 26, 11, pp. 832--843 (1983)
- 16. Ramez, E.: Fundamentals of database systems, Pearson Education, 4th ed., pp. 842--856 (2004)